the globus alliance
www.globus.org

Reliable Data Movement Framework for Distributed Science Environments

Raj Kettimuthu
Argonne National Laboratory and
The University of Chicago

the globus alliance
www.globus.org

Outline

- Introduction
- Motivation
- Data Transfer Problem
- Requirements
- Reliable Data Movement Framework
- Future Directions

Today's Science Environments

- Science environment today is very different
- Large-scale collaborative science is becoming increasingly common
- Need for distributed community of users to access and analyze large amounts of data reliably is a fundamental requirement
- This requirement arises in both simulation and experimental sciences

Simulation Science

- In simulation science, the data sources are supercomputer simulations
  - For eg, DOE-funded climate modeling groups generate large reference simulations at supercomputer centers
  - Many climate scientists need to extract and analyze subsets of this data in various ways
- Combustion, fusion, computational chemistry, and astrophysics communities have similar requirements for remote and distributed data analysis

Experimental Science

- Data sources are facilities such as high energy and nuclear physics experiments and light sources.
  - For eg, the experimental program based upon the LHC at the CERN will produce petabytes of raw data per year for approximately 15 years
  - Thousands of physicists worldwide will participate in the production and analysis of simulated and derived data sets from this raw experimental data
- DOE light sources can also produce large quantities of data that must be distributed, analyzed, and visualized
- The international fusion experiment, ITER

Science Environments

- Raw simulation or observational data is just a starting point for most investigations
- Understanding comes from further analysis, reduction, visualization, and exploration
- Analysis must often be performed on a different class of petascale resource, a smaller resource such as a cluster, or even a scientist's desktop
- Furthermore the data is a community asset that must be accessible to any member of a distributed collaboration

Network Capabilities

- Scientist A is in California
- Scientist B is in New York
- They both are connected through the Internet
- Scientist A wants to transfer 1 Terabyte of data to Scientist B
- What is the fastest way to transfer the data?

FedEx

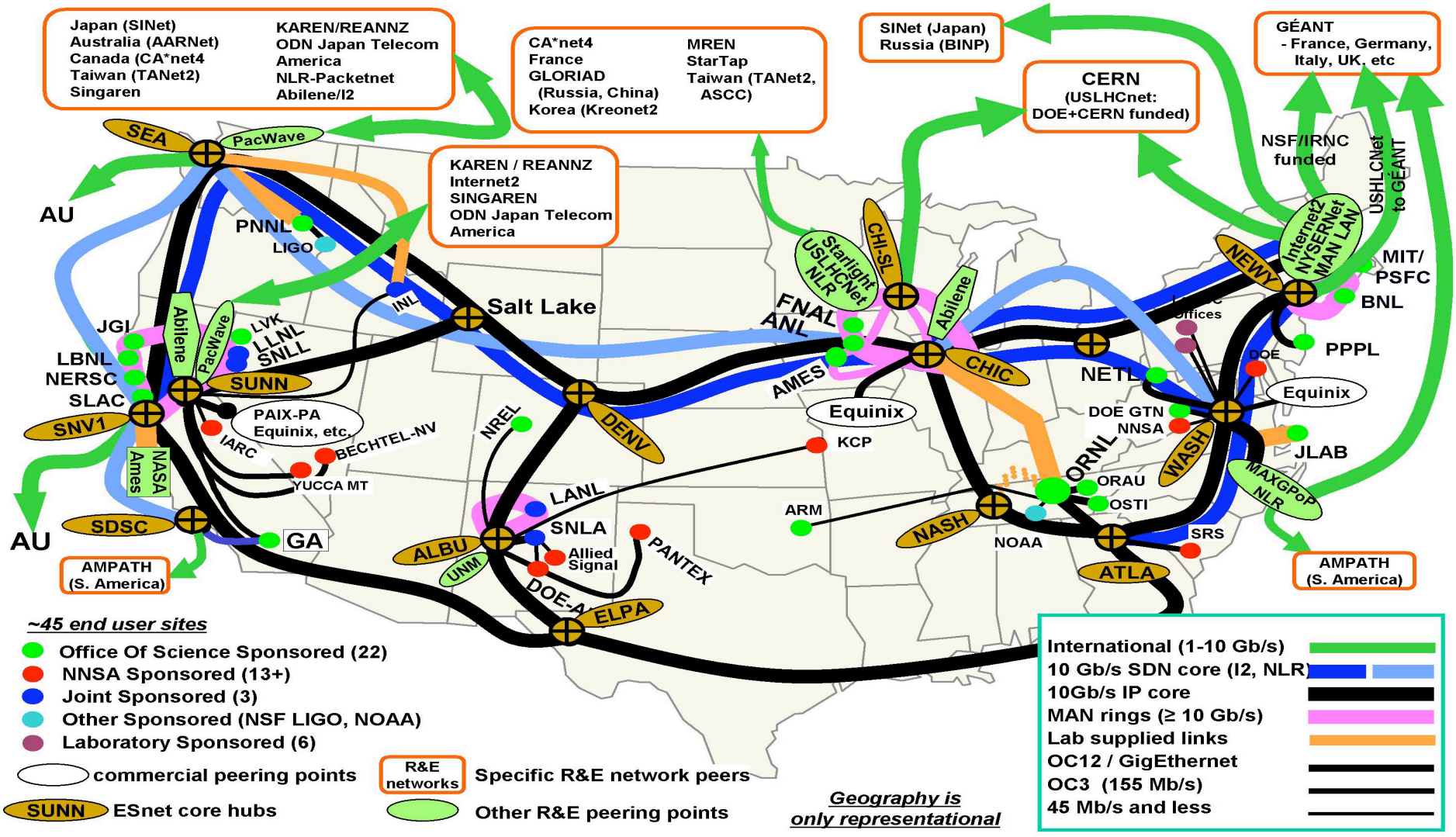
Network Capabilities

- Until a few years ago, Tri-labs (Los Alamos, Lawrence Livermore and Sandia) transferred data via tapes sent thru fedex
- To transfer 100 TB in 24 hours, need a sustained data rate > 9.5 Gbit/s
- 10 Gbit/s networks are becoming increasingly common in scientific environments
  - DOE's ESNet, UltraScience Net, Science Data Networks and Internet2 has 10Gb/s or higher links
  - Thanks to the advancement in networking technologies

the globus alliance
www.globus.org

ESNET

ESnet Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (12/2007)



End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross the US in less than a second

Yes - But it would still take two hours to get downtown!

End-to-end problem

- Data movement in distributed science environments is an end-to-end problem
- A 10 Gbit/s network link between the source and destination does not guarantee an end-to-end data rate of 10 Gbit/s
- Other factors such as storage system, disk, data rate supported by the end node
- Deal with failures of various sorts
  - Firewalls can cause difficulties

End-to-end data transfer

- Efficient and robust wide area data transport requires the management of complex systems at multiple levels.
- For example, in a recent work, we required 32 hosts connected at 1 Gbit/s to drive a 30 Gbit/s connection.
- Effective end-to-end data transfers thus demand a systems approach
  - Integrates file systems, computers, network interfaces, and network protocols
  - Encapsulated in easily usable and portable software

Requirements

- Fast
- Secure
- Reliable
- Extensible
- Standard
- Robust

the globus alliance
www.globus.org

GridFTP

- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol - defines extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- GridFTP is the OGF recommended data movement protocol

GridFTP

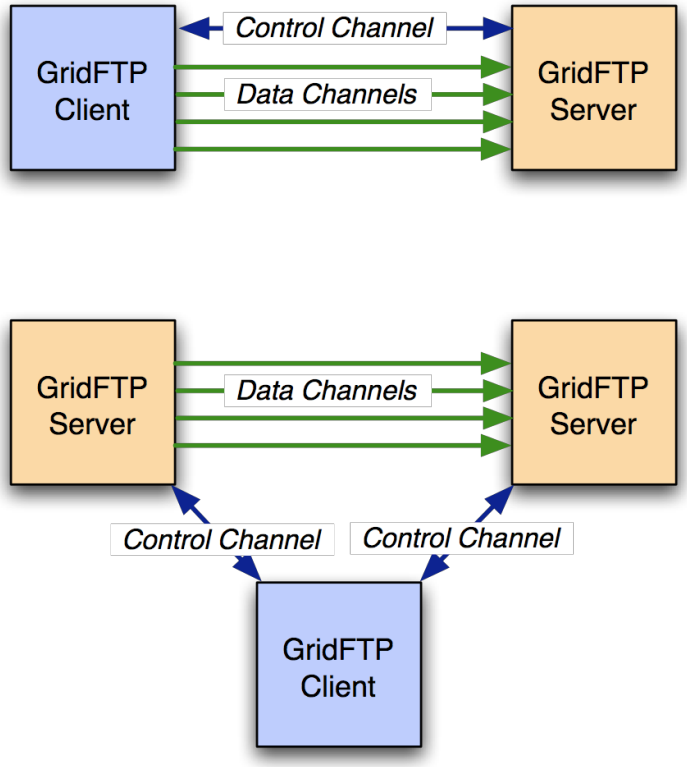
- We (Globus Alliance) supply a reference implementation:
  - Server
  - Client tools
  - Development Libraries
- Multiple independent implementations can interoperate
  - Fermi Lab and U. Virginia have home grown servers that work with ours

Requirements

- Fast
- Secure
- Reliable
- Extensible
- Standard ✓
- Robust

GridFTP

- Two channel protocol like FTP
- Control Channel
  - Communication link (TCP) over which commands and responses flow
  - Low bandwidth; encrypted and integrity protected by default
- Data Channel
  - Communication link(s) over which the actual data of interest flows
  - High Bandwidth; authenticated by default; encryption and integrity protection optional

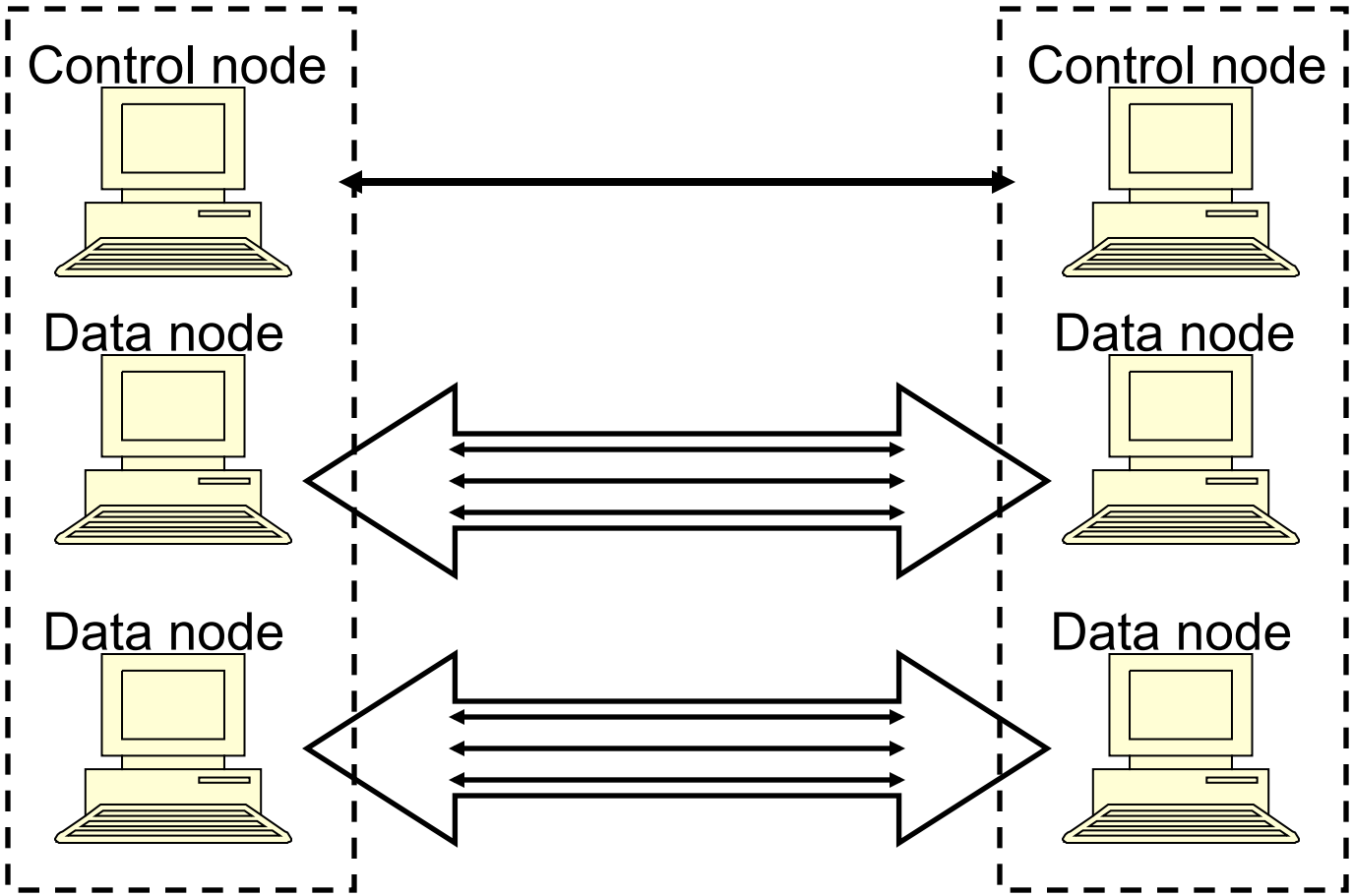


Globus GridFTP Features

- GridFTP is Fast
  - Parallel TCP streams
  - Non TCP protocol such as UDT
  - Set TCP buffer sizes
  - Order of magnitude greater
- Cluster-to-cluster data movement
  - Co-ordinated data movement using multiple computers at each end
  - Another order of magnitude

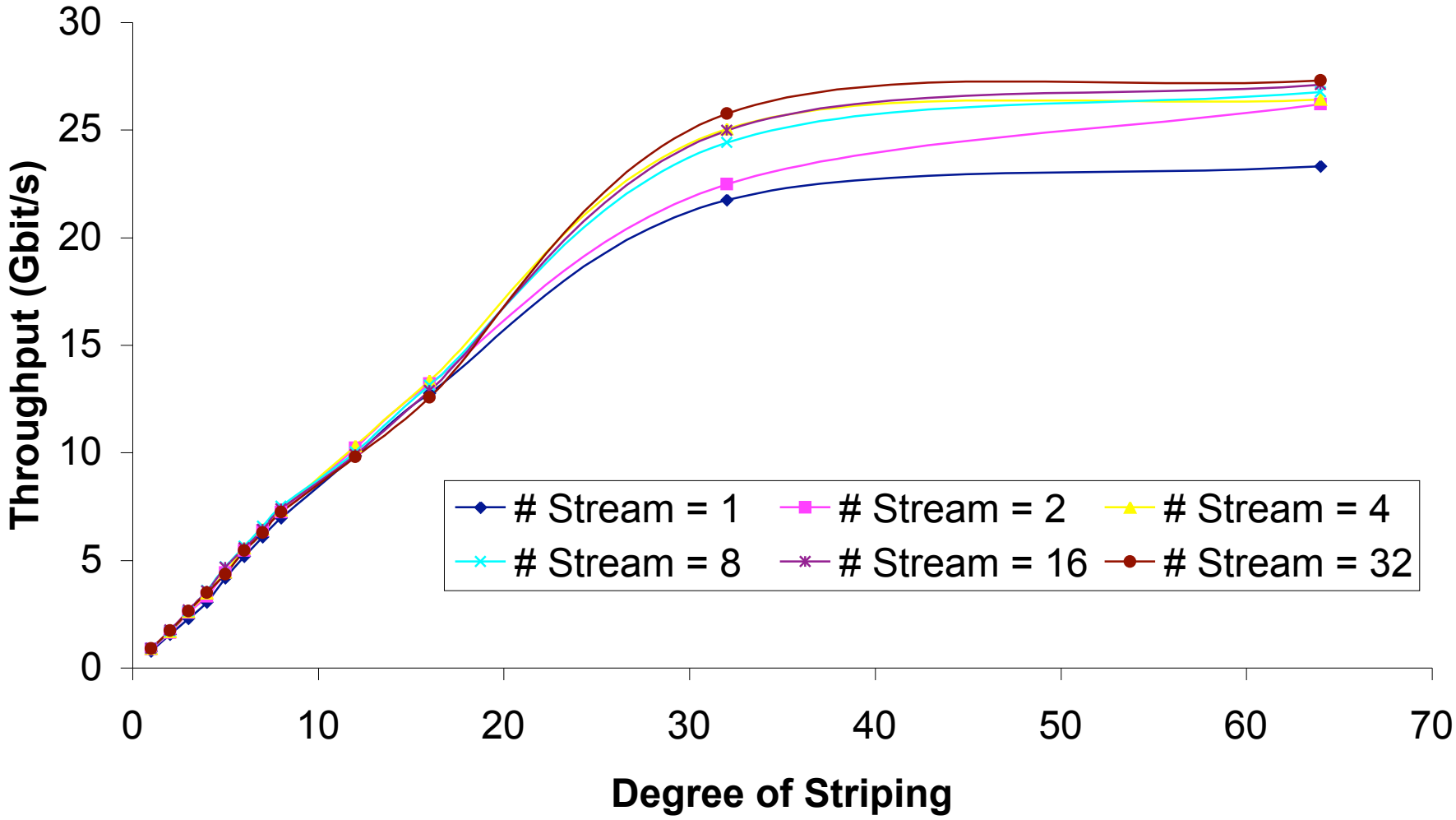
the globus alliance
www.globus.org

Cluster-to-Cluster transfers



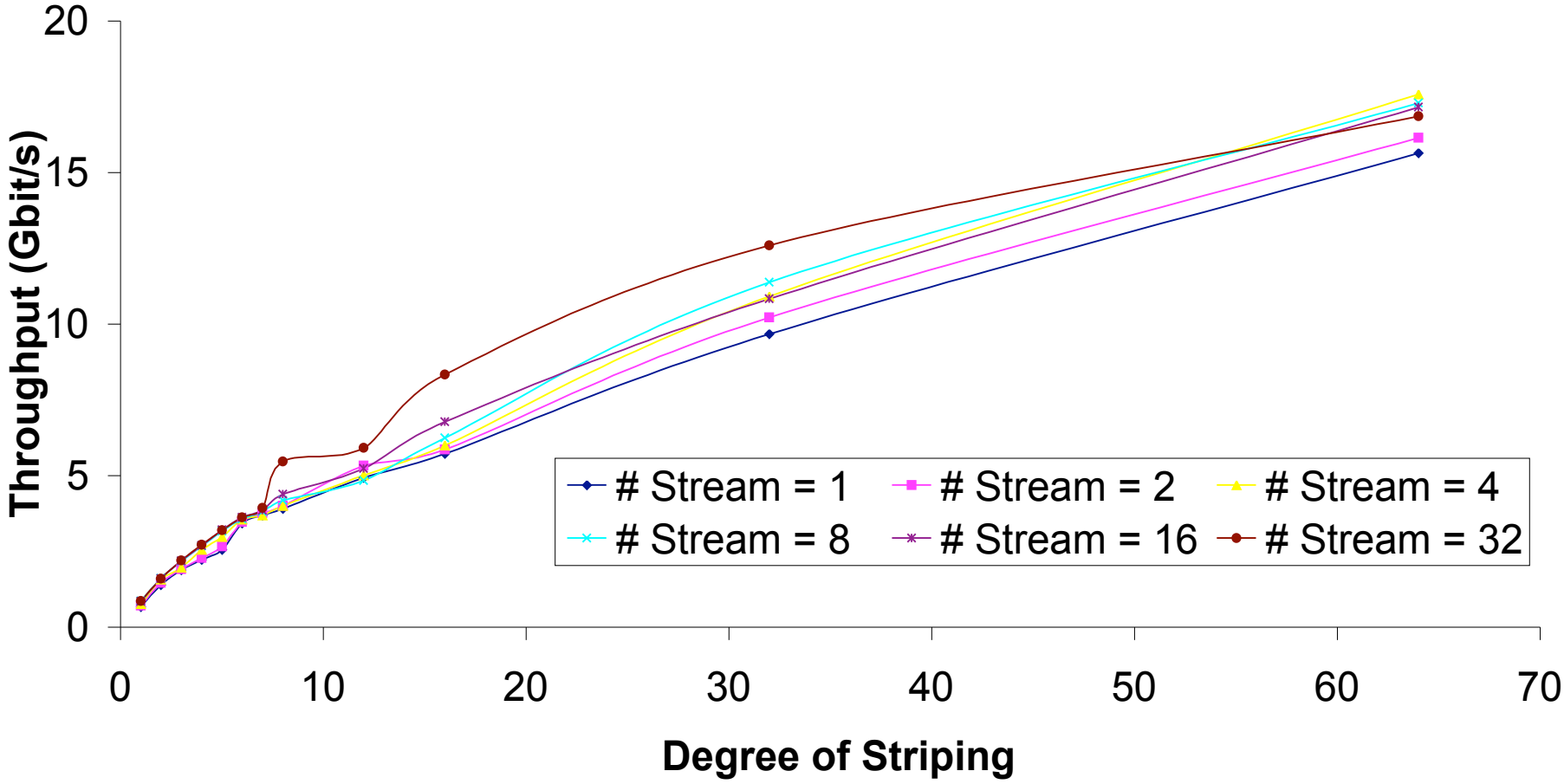
Performance

- Mem. transfer between Urbana, IL and San Diego, CA



Performance

- Disk transfer between Urbana, IL and San Diego, CA



the globus alliance
www.globus.org

Requirements

- Fast ✓
- Secure
- Reliable
- Extensible
- Standard ✓
- Robust

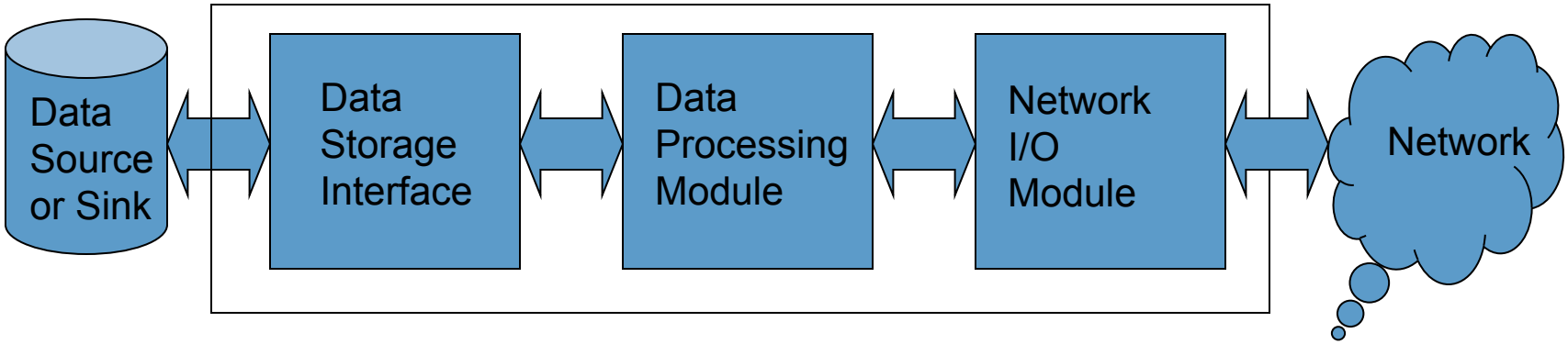
Security

- Often there is need to authenticate clients and control access to the data
- Globus GridFTP supports multiple security mechanisms to authenticate and authorize clients
  - Anonymous access
  - Username/password
  - SSH security
  - Grid Security Infrastructure (GSI)

Requirements

- Fast ✓
- Secure ✓
- Reliable
- Extensible
- Standard ✓
- Robust

Modular



Well defined interfaces

Data Storage Interface

- POSIX file system
- High Performance Storage System (HPSS)
- Storage Resource Broker (SRB)
- Freeloader (under development)

Modular

- Network I/O module
  - TCP
  - Easy to plug-in external libraries
  - UDT
  - Phoebus
- Data processing module
  - Compression (under development)
  - Checksum

Requirements

- Fast ✓
- Secure ✓
- Reliable
- Extensible ✓
- Standard ✓
- Robust

GridFTP in production

- GridFTP has been around for many years
- Many Scientific communities rely on GridFTP
  - HEP community is basing its entire tiered data movement infrastructure for the LHC computing Grid on GridFTP
  - Southern California Earthquake Center (SCEC), Laser Interferometer Gravitational-Wave Observatory (LIGO), Earth Systems Grid (ESG), Relativistic Heavy Ion Collider (RHIC), Advanced Photon Source use GridFTP for data movement
  - European Space Agency, Disaster Recovery Center in Japan, British Broadcasting Corporation move large volumes of data using GridFTP
- GridFTP facilitates an average of more than 3 million data transfers every day

Requirements

- Fast ✓
- Secure ✓
- Reliable
- Extensible ✓
- Standard ✓
- Robust ✓

Handling failures

- GridFTP server sends restart and performance markers periodically
  - Default every 5s - configurable
- Helpful if there is any failure
  - No need to transfer the entire file again
  - Can start from the last restart marker
- GridFTP supports partial file transfers

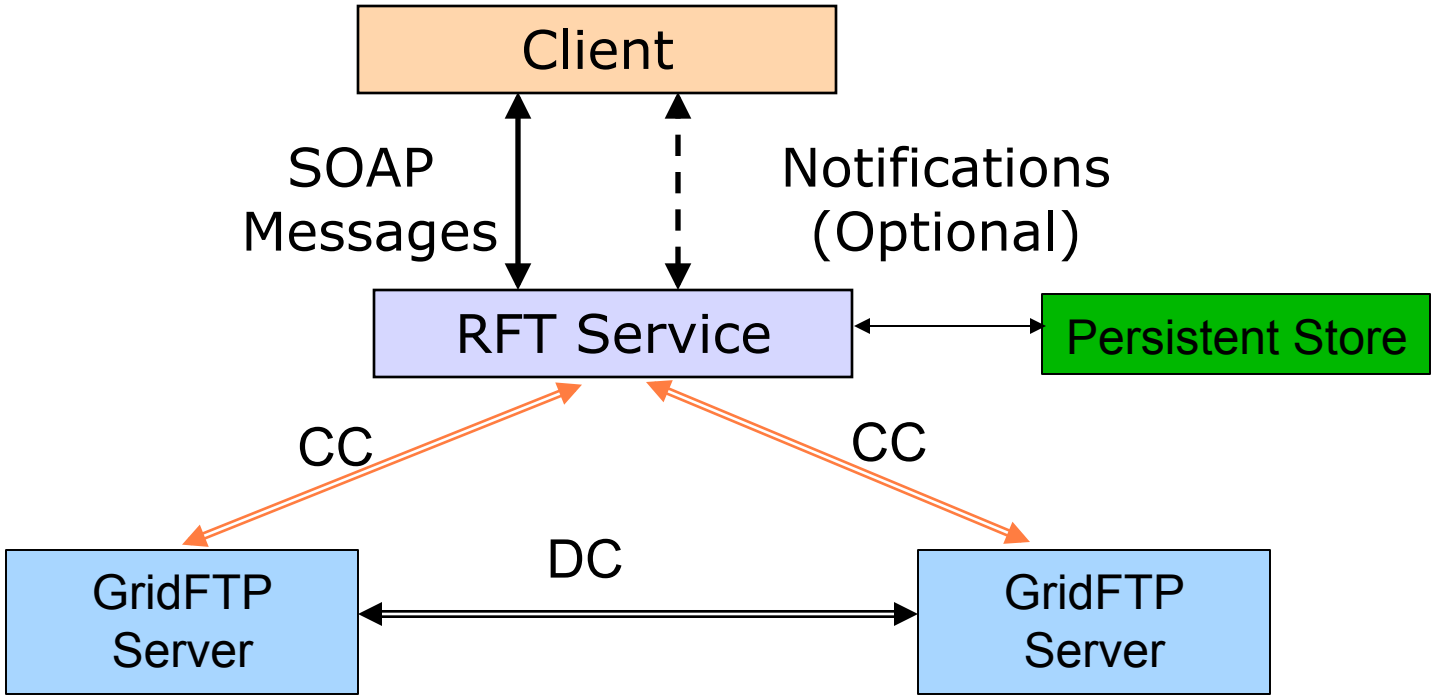
GridFTP clients

- Globus-url-copy - commonly used command-line client
- Lots of people have developed clients independent of the Globus Project
  - Uberftp
- These clients support transfer retries and recover from server failures
- What if the client fails in the middle of a transfer?

Globus Reliable File Transfer Service (RFT)

- GridFTP client that provides more reliability
- GridFTP - on demand transfer service
  - Not a queuing service
- RFT
  - Queues requests
  - Orchestrates transfers on client's behalf
  - Writes to persistent store
  - Recovers from GridFTP and RFT service failures

RFT


Client
SOAP
Messages
Notifications
(Optional)
RFT Service
Persistent Store
CC
CC
DC
GridFTP
Server
GridFTP
Server

Requirements

- Fast ✓
- Secure ✓
- Reliable ✓
- Extensible ✓
- Standard ✓
- Robust ✓

GridFTP

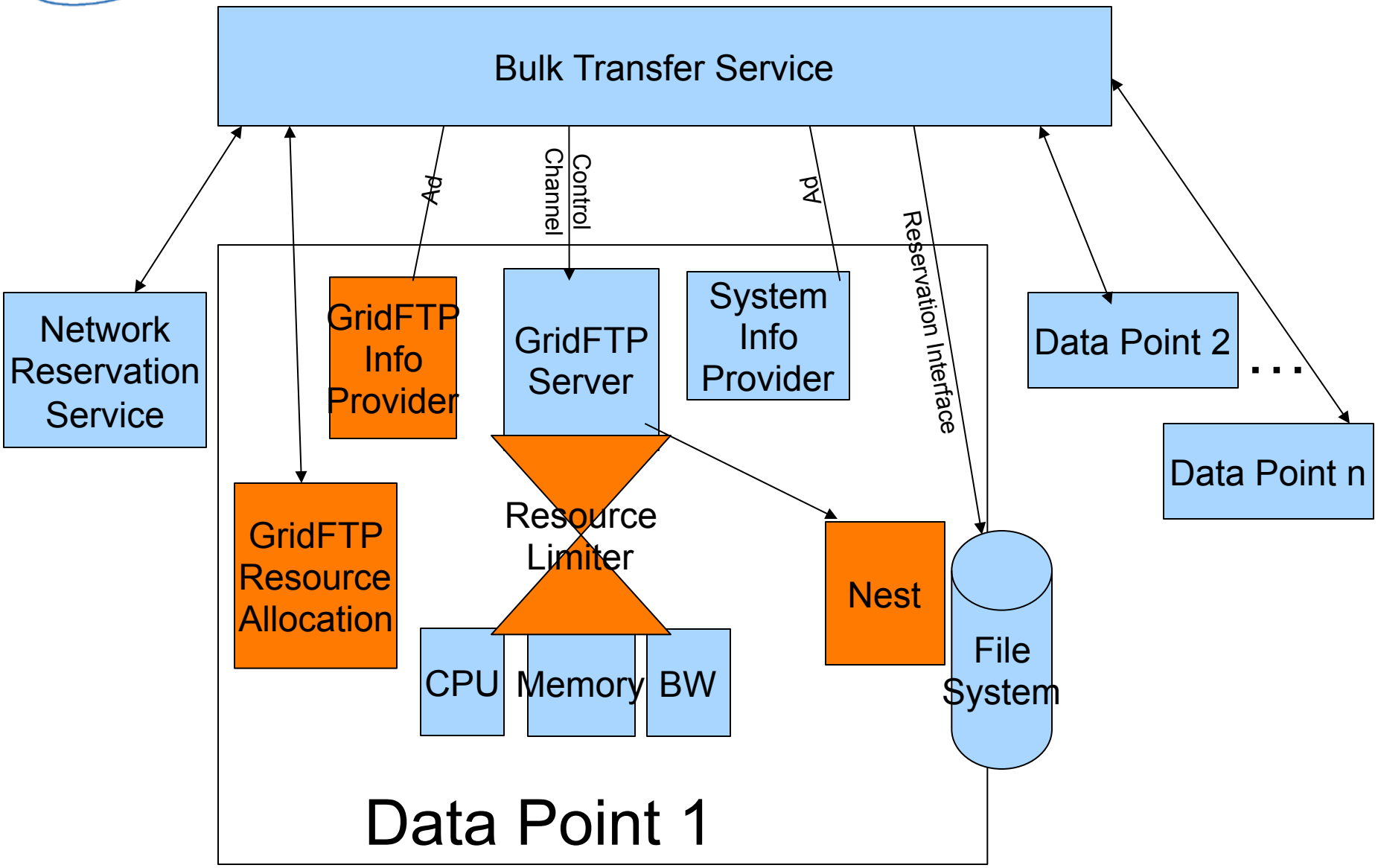
Best effort service

- Data movement in distributed environments is still on best effort basis
- No Quality of Service (QoS) guarantees
- Network is shared
- Limited disk space
  - Destination might run out of space in the middle of a transfer
- End node, network, disk can fail any time

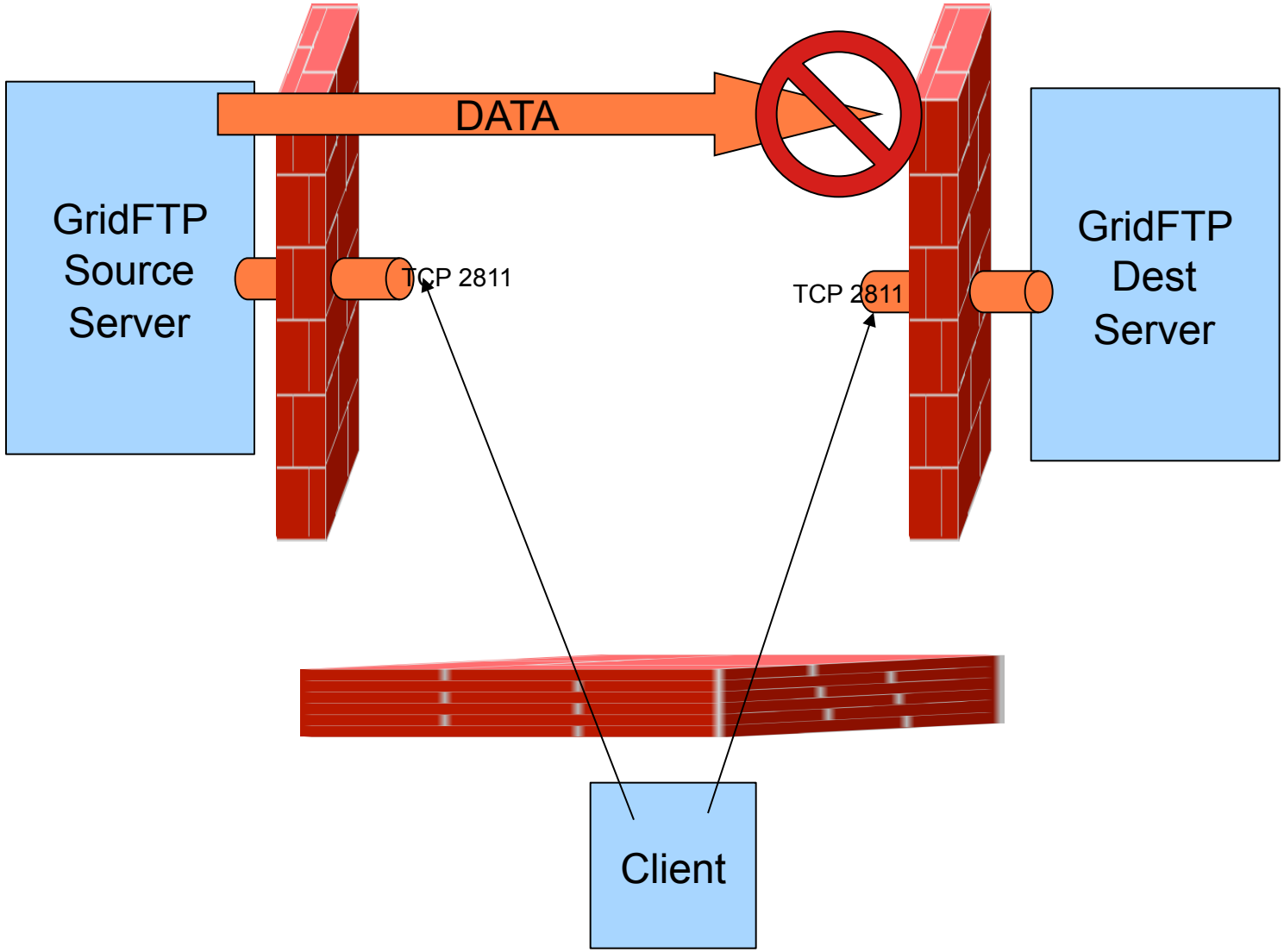
Better than best effort

- Advances in network and storage reservations
  - Internet2 Dynamic Circuits Network
  - ESNet OSCARS
  - DOE sponsored LambdaStation and TeraPaths
  - Reserve bandwidth on the network
  - Storage Reservation Managers (SRM), NeST allows to reserve disk space

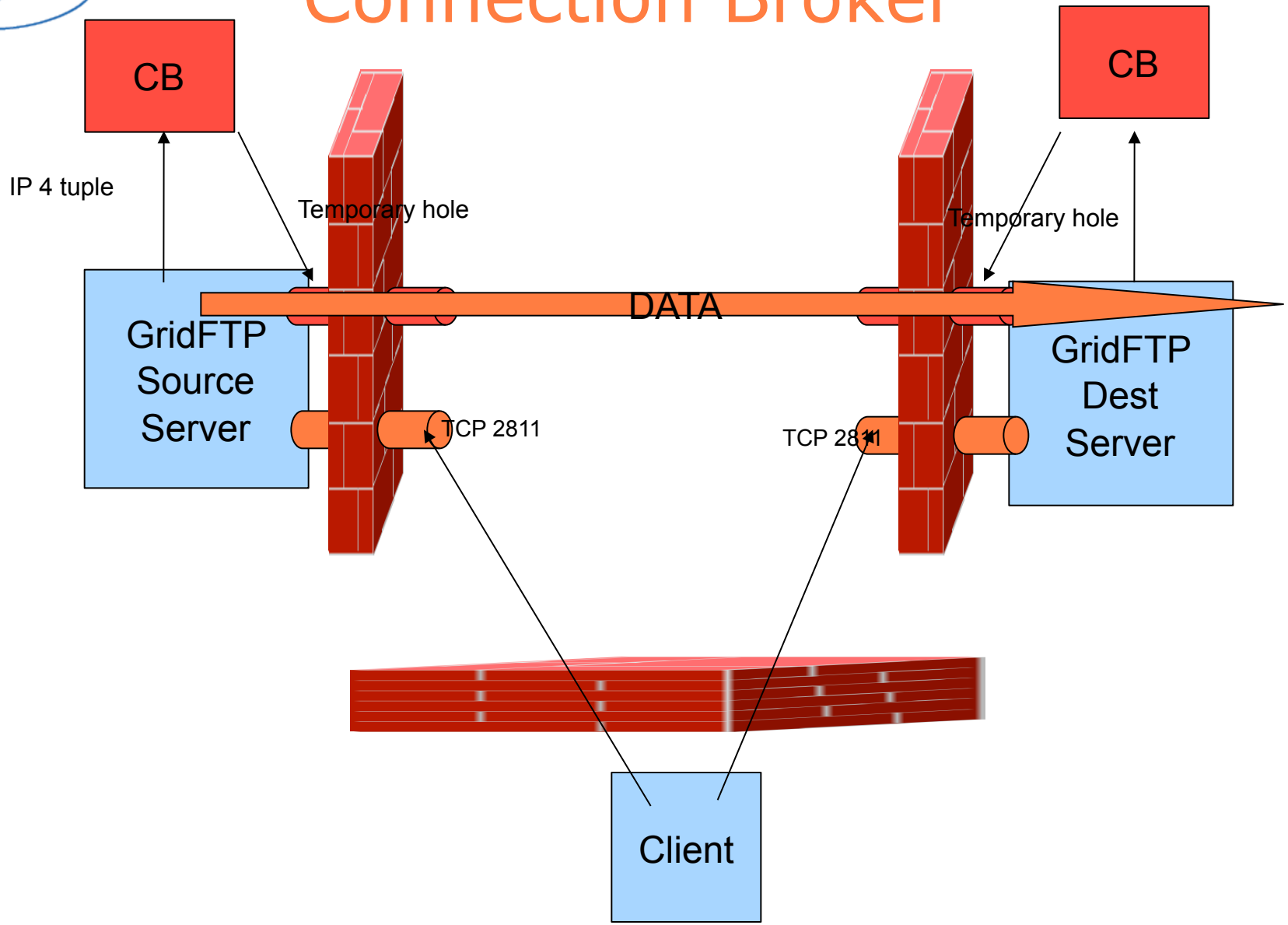
Better than best effort



Firewall


DATA
GridFTP Source Server
GridFTP Dest Server
TCP 2811
TCP 2811
Client

Connection Broker



the globus alliance
www.globus.org

Links and contacts

- GridFTP is available in the Globus toolkit
- Latest version available at http://www.globus.org/toolkit/downloads/4.2.0/
- Documentation available at http://www.globus.org/toolkit/docs/4.2/4.2.0/data/gridftp/index.html
- Simple to install
  - Configure; make gridftp install;
  - Installs only gridftp and its dependencies
  - Binaries available for many platforms
- Gridftp-user@globus.org, gridftp-dev@globus.org
- Kettimut@mcs.anl.gov

Questions